1

Encoding audio signals

FIELD OF THE INVENTION

The invention relates to an encoder for audio signals, and a method of encoding audio signals.

5    BACKGROUND OF THE INVENTION

Within the field of audio coding it is generally desired to encode an audio signal in order to reduce the bit rate without unduly compromising the perceptual quality of the audio signal. The reduced bit rate is advantageous for limiting the bandwidth when communicating the audio signal or the amount of storage required for storing the audio

10    signal.

Parametric descriptions of audio signals have gained interest during the last years, especially in the field of audio coding. It has been shown that transmitting (quantized) parameters which describe audio signals require only a limited transmission capacity to enable to synthesize perceptually substantially equal audio signals at the receiving end.

15    US2003/0026441 discloses the synthesizing of an auditory scene by applying two or more different sets of one or more spatial parameters (e.g. an inter-ear level difference ILD, or an inter-ear time difference ITD) to two or more different frequency bands of a combined audio signal, wherein each different frequency band is treated as if it corresponds to a single audio source in the auditory scene. In one embodiment, the combined audio signal

20    corresponds to the combination of the left and right audio signals of a binaural signal corresponding to an input auditory scene. The different sets of spatial parameters are applied to reconstruct the input auditory scene. The transmission bandwidth requirements are reduced by reducing to one the number of different audio signals that need to be transmitted to a receiver configured to synthesize/reconstruct the auditory scene.

25    In the transmitter, a TF transform is applied to corresponding parts of each of the left and right audio signals of the input binaural signal to convert the signals to the frequency domain. An auditory scene analyzer processes the converted left and right audio signals in the frequency domain to generate a set of auditory scene parameters for each one of a plurality of different frequency bands in those converted signals. For each corresponding

2

pair of frequency bands, the analyzer compares the converted left and right audio signals to generate one or more spatial parameters. In particular, for each frequency band, the cross-correlation function between the converted left and right audio signals is estimated. The maximum value of the cross-correlation indicates how much the two signals are correlated.

5    The location in time of the maximum of the cross-correlation corresponds to the ITD. The ILD can be obtained by computing the level difference of the power values of the left and right audio signals.

SUMMARY OF THE INVENTION

10   It is an object of the invention to provide an encoder for encoding audio signals which requires less processing power.

To reach this object, a first aspect of the invention provides an encoder for encoding audio signals. A second aspect of the invention provides a method of encoding audio signals. Advantageous embodiments are defined in the dependent claims.

15   The encoder disclosed in US2003/0026441 first transforms the audio signals from the time domain to the frequency domain. This transformation is usually referred to as the Fast Fourier Transform, further referred to as FFT. Usually, the audio signal in the time domain is divided into a sequence of time segments or frames, and the transformation to the frequency domain is performed sequentially for each one of the frames. The relevant part of

20   the frequency domain is divided into frequency bands. In each frequency band the cross-correlation function is determined of the input audio signals. This cross-correlation function has to be transformed from the frequency domain to the time domain. This transformation is usually referred to as the inverse FFT further referred to as IFFT. In the time domain, the maximum value of the cross-correlation function has to be determined to find the location in

25   time of this maximum and thus the value of the ITD.

The encoder in accordance with the first aspect of the invention also has to transform the audio signals from the time domain to the frequency domain, and also has to . determine the cross-correlation function in the frequency domain. In the encoder in accordance with the invention, the spatial parameter used is the inter-channel phase

30   difference further referred to as IPD or the inter-channel coherence further referred to as IC, or both. Also other spatial parameters such as the inter-channel level differences further referred to as ILD may be coded. The inter-channel phase difference IPD is comparable with the inter-ear time difference ITD of the prior art.

3

However instead of performing the IFFT and the search for the maximum value of the cross-correlation function in the time domain, a complex coherence value is calculated by summing the (complex) cross-correlation function values in the frequency domain. The inter-channel phase difference IPD is estimated by the argument of the complex

5    coherence value, the inter-channel coherence IC is estimated by the absolute value of the complex coherence value.

In the prior art US2003/0026441, the inverse FFT and the search for the maximum of the cross-correlation function in the time domain requires a high amount of processing effort. This prior art is silent about the determination of the coherence parameter.

10    In the encoder in accordance with the invention the inverse FFT is not required, the complex coherence value is calculated by summing the (complex) cross-correlation function values in the frequency domain. Either the IPD or the IC, or the IPD and the IC are determined in a simple manner from this sum. Thus, the high computational effort for the inverse FFT is replaced by a simple summing operation. Consequently, the approach

15    in accordance with the invention requires less computational effort.

It should be noted that although prior art US2003/0026441 uses an FFT to yield a complex-valued frequency-domain representation of the input signals, complex filter banks may also be used. Such filter banks use complex modulators to obtain a set of band-limited complex signals (cf. Ekstrand, P. (2002). Bandwidth extension of audio signals by

20    spectral band replication. Proc. 1$^{st}$ Benelux Workshop on model based processing and coding of audio (MPCA-2002), Leuven, Belgium). The IPD and IC parameters can be computed in a similar way as for the FFT, with the only difference that summation is required across time instead of frequency bin.

In an embodiment as defined in claim 2, the cross-correlation function is

25    calculated as a multiplication of one of the input audio signals in a band-limited, complex domain and the complex conjugated other one of the input audio signals to obtain a complex cross-correlation function which can be thought to be represented by an absolute value and an argument.

In an embodiment as defined in claim 3, a corrected cross-correlation function

30    is calculated as the cross-correlation function wherein the argument is replaced by the derivative of said argument. At high frequencies, it is known that the human auditory system is not sensitive to fine-structure phase-differences between the two input channels. However, considerable sensitivity to the time difference and coherence of the envelope exists. Hence at high frequencies, it is more relevant to compute the envelope ITD and envelope coherence

4

for each frequency band. However, this requires an additional step of computing the (Hilbert) envelope. In the embodiment in accordance with the invention as defined in claim 3, it is possible to calculate the complex coherence value by summing the corrected cross-correlation function directly in the frequency domain. Again, the IPD and/or IC can be
5    determined in a simple manner from this sum as the argument and phase of the sum, respectively.

In an embodiment as defined in claim 4, the frequency domain is divided into a predetermined number of frequency sub-bands, further also referred to as sub-bands. The frequency range covered by different sub-bands may increase with the frequency. The
10   complex cross-correlation function is determined for each sub-band, by using both the input audio signals in the frequency domain in this sub-band. The input audio signals in the frequency domain in a particular one of the sub-bands are also referred to as sub-band audio signals. The result is a cross-correlation function for each one of the sub-bands. Alternatively, the cross-correlation function may only be determined for a sub-set of the sub-bands,
15   depending on the required quality of the synthesized audio signals. The complex coherence value is calculated by summing the (complex) cross-correlation function values in each of the sub-bands. And thus, also the IPD and/or IC are determined per sub-band. This sub-band approach enables to provide a different coding for different frequency sub-bands and allows to further optimize the quality of the decoded audio signal versus the bit-rate of the coded
20   audio signal.

In an embodiment as defined in claim 5, for lower frequencies, the complex cross-correlation functions per sub-band are obtained by multiplying one of the sub-band audio signals with the complex conjugated other one of the sub-band audio signals. The complex cross-correlation function has an absolute value and an argument. The complex
25   coherence value is obtained by summing the values of the cross-correlation function in each of the sub-bands. For higher frequencies, corrected cross-correlation functions are determined which are determined in the same manner as the cross-correlation functions for lower frequencies but wherein the argument is replaced by a derivative of this argument. Now, the complex coherence value per sub-band is obtained by summing the values of the corrected
30   cross-correlation function per sub-band. The IPD and/or IC are determined in the same manner from the complex coherence value, independent on the frequency.

These and other aspects of the invention are apparent from and will be elucidated with reference to the embodiments described hereinafter.

5

BRIEF DESCRIPTION OF THE DRAWINGS

In the drawings:

Fig. 1 shows a block diagram of an audio encoder,

Fig. 2 shows a block diagram of an audio encoder of an embodiment in

5    accordance with the invention,

Fig. 3 shows a block diagram of part of the audio encoder of another

embodiment in accordance with the invention, and

Fig. 4 shows a schematic representation of the sub-band division of the audio

signals in the frequency domain.

10

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Fig. 1 shows a block diagram of an audio encoder. The audio encoder receives

two input audio signals $x(n)$ and $y(n)$ which are digitized representations of, for example, the

left audio signal and the right audio signal of a stereo signal in the time domain. The indices

15    n refer to the samples of the input audio signals $x(n)$ and $y(n)$. The combining circuit 1

combines these two input audio signals $x(n)$ and $y(n)$ into a monaural signal MAS. The stereo

information in the input audio signals $x(n)$ and $y(n)$ is parameterized in the parameterizing

circuit 10 which comprises the circuits 100 to 113 and supplies, by way of example only, the

parameters ITDi, the inter-channel time difference per frequency sub-band (or the IPDi:

20    inter-channel phase difference per frequency sub-band) and CIi (inter-channel coherence per

frequency sub-band). The monaural signal MAS and the parameters ITDi, ICi are transmitted

in a transmission system or stored on a storage medium (not shown). At the receiver or

decoder (not shown) the original signals $x(n)$ and $y(n)$ are reconstructed from the monaural

signal MAS and the parameters ITDi, ICi.

25    Usually, the input audio signals $x(n)$ and $y(n)$ are processed per time segment

or frame. The segmentation circuit 100 receives the input audio signal $x(n)$ and stores the

received samples during a frame to be able to supply the stored samples $Sx(n)$ of the frame to

the FFT-circuit 102. The segmentation circuit 101 receives the input audio signal $y(n)$ and

stores the received samples during a frame to be able to supply the stored samples $Sy(n)$ of

30    the frame to the FFT-circuit 103.

The FFT-circuit 102 performs a Fast Fourier Transformation on the stored

samples $Sx(n)$ to obtain an  audio signal $X(k)$ in the frequency domain. In the same manner,

the FFT-circuit 103 performs a Fast Fourier Transformation on the stored samples $Sy(n)$ to

obtain an audio signal $Y(k)$ in the frequency domain. The sub-band dividers 104 and 105

6

receive the audio signals X(k) and Y(k), respectively, to divide the frequency spectra of these audio signals X(k) and Y(k) into frequency sub-bands i (see Fig. 4) to obtain the sub-band audio signals Xi(k) and Yi(k). This operation is further elucidated with respect to Fig. 4.

The cross-correlation determining circuit 106 calculates the complex cross-

5    correlation function Ri of the sub-band audio signals Xi(k) and Yi(k) for each relevant sub-band. Usually, the cross-correlation function Ri is obtained in each relevant sub-band by multiplying one of the audio signals in the frequency domain Xi(k) with the complex conjugated other one of the audio signals in the frequency domain Yi(k). It would be more correct to indicate the cross-correlation function with Ri(X,Y)(k) or Ri(X(k),Y(k)), but for

10   clarity this is abbreviated to Ri.

The optional normalizing circuit 107 normalizes the cross-correlation function Ri to obtain a normalized cross-correlation function Pi(X,Y)(k) or Pi(X(k),Y(k)) which is abbreviated to Pi:


15              Pi = Ri(Xi,Yi) / sqrt( sum (Xi(k) . conj Xi(k)) * (sum Xi(k) . conj Xi(k)))


wherein sqrt is the square root, and conj is the complex conjugation.

It is to be noted that this normalization process requires the computation of the energies of the sub-band signals Xi(k), Yi(k) of the two input signals x(n), y(n). However, this operation

20   is required anyway in order to compute the inter-channel intensity difference IID for the current sub-band i. The IID is determined by the quotient of these energies. Thus, the cross function Ri can be normalized by taking the goniometric mean of the corresponding sub-band intensities of the two input signals Xi(k), Yi(k).

The known IFFT (Inverse Fast Fourier Transform) circuit 108 transforms the

25   normalized cross-correlation function Pi in the frequency domain back to the time domain, yielding the normalized cross-correlation ri(x(n),y(n)) or ri(x,y)(n) in the time domain which is abbreviated as ri. The circuit 109 determines the peak value of the normalized cross-correlation ri. The inter-channel time delay ITDi for a particular sub-band is the argument n of the normalized cross-correlation ri at which the peak value occurs. Or said in other words,

30   the delay which corresponds to this maximum in the normalized cross-correlation ri is the ITDi. The inter-channel coherence ICi for the particular sub-band is the peak value. The ITDi provides the required shift of the two input audio signals x(n), y(n) with respect to each other to obtain the highest possible similarity. The ICi indicates how similar the shifted input audio

7

signals x(n), y(n) are in each sub-band. Alternatively, the IFFT may be performed on the not normalized cross-correlation function Ri.

Although this block diagram shows separate blocks performing operations, the operations may be performed by a single dedicated circuit or integrated circuit. It is also
5    possible to perform all the operations or a part of the operations by a suitably programmed microprocessor.

Fig. 2 shows a block diagram of an audio encoder of an embodiment in accordance with the invention. This audio encoder comprises the same circuits 1, and 100 to 107 as shown in Fig. 1 which operate in the same manner. Again, the optional normalizing
10   circuit 107 normalizes the cross-correlation function Ri to obtain a normalized cross-correlation function Pi. The coherence value computing circuit 111 computes a complex coherence value Qi for each relevant sub-band i by summing the complex normalized cross-correlation function Pi:

15   $$Qi = \text{sum } (Pi(Xi(k),Yi(k)))$$

The FFT-bin index k is determined by the bandwidth of each sub-band. Preferably, to minimize computation efforts, only the positive (k = 0 to K/2, where K is the FFT size) or negative frequencies (k = -K/2 to 0) are summed. This computation is performed in the
20   frequency domain and thus does not require an IFFT to first transform the normalized cross-correlation function Pi to the time domain. The coherence estimator 112 estimates the coherence ICi with the absolute value of the complex coherence value Qi. The phase difference estimator 113 estimates the IPDi with the argument or angle of the complex coherence value Qi.

25   Thus now, the inter-channel coherence ICi and the inter-channel phase difference IPDi are obtained for each relevant sub-band i without requiring, in each relevant sub-band, an IFFT operation and a search for the maximum value of the normalized cross-correlation ri. This saves a considerable amount of processing power. Alternatively, the complex coherence value Qi may be obtained by summing the not normalized cross-
30   correlation function Ri.

Fig. 3 shows a block diagram of part of the audio encoder of another embodiment in accordance with the invention.

For high frequencies, for example above 2 kHz or above 4 kHz, in the prior art (cf. Baumgarte, F., Faller. C (2002). Estimation of auditory spatial cues for binaural cue

8

coding. Proc. ICASSP'02), the envelope coherence may be calculated which is even more
computational intensive than computing the waveform coherence as elucidated with respect
to Fig. 1. Experimental results demonstrated that the envelope coherence can be fairly
accurately estimated by replacing the phase values ARG of the frequency domain
5    (normalized) complex cross-correlation function Ri by the derivative DA of these phase
values ARG.

Fig. 3 shows the same cross-correlation determining circuit 106 as in Fig. 1.
The cross-correlation determining circuit 106 calculates the complex cross-correlation
function Ri of the sub-band audio signals $Xi(k)$ and $Yi(k)$ for each relevant sub-band.
10   Usually, the cross-correlation function Ri is obtained in each relevant sub-band by
multiplying one of the audio signals in the frequency domain $Xi(k)$ with the complex
conjugated other one of the audio signals in the frequency domain $Yi(k)$. The circuit 114
which receives the cross-correlation function Ri comprises a calculation unit 1140 which
determines the derivative DA of the argument ARG of this complex cross-correlation
15   function Ri. The amplitude AV of the cross-correlation function Ri is unchanged. The output
signal of the circuit 114 is a corrected cross-correlation function $R'i(Xi(k),Yi(k))$ (which is
also referred to as R'i) which has the amplitude AV of the cross-correlation function Ri and
an argument which is the derivative DA of the argument ARG:

20
$$| R'i(Xi(k),Yi(k)) | = | Ri(Xi(k),Yi(k)) | \text{ and}$$

$$arg(R'i(Xi(k),Yi(k))) = d(arg(Ri(Xi(k),Yi(k))))/dk$$

The coherence value computing circuit 111 computes a complex coherence value Qi for each
25   relevant sub-band i by summing the complex cross-correlation function R'i. Thus, instead of
the computational intensive Hilbert envelope approach now only simple operations are
required.

The above described approach can of course also be applied on the normalized
complex cross-correlation function Pi to obtain a corrected complex normalized cross-
30   correlation function P'i.

Fig. 4 shows a schematic representation of the sub-band division of the audio
signals in the frequency domain. Fig. 4A shows how the audio signal $X(k)$ in the frequency
domain is divided into sub-band audio signals $Xi(k)$ in sub-bands i of the frequency spectrum
f. Fig. 4B shows how the audio signal $Y(k)$ in the frequency domain is divided into sub-band

audio signals Yi(k) in sub-bands i of the frequency spectrum f. The frequency-domain signals X(k) and Y(k) are grouped into sub-bands i, resulting in sub-bands Xi(k) and Yi(k). Each sub-band Xi(k) corresponds to a certain range of FFT-bin indexes k=[ksi...kei], where ksi and kei indicate the first and last FFT bin index k, respectively. Similarly each subband

5    Yi(k) corresponds to the same range of FFT-bin indexes k.

It should be noted that the above-mentioned embodiments illustrate rather than limit the invention, and that those skilled in the art will be able to design many alternative embodiments without departing from the scope of the appended claims.

The invention is not limited to stereo signals and may, for example, be

10    implemented on multi-channel audio as used in DVD and SACD.

In the claims, any reference signs placed between parentheses shall not be construed as limiting the claim. Use of the verb "comprise" and its conjugations does not exclude the presence of elements or steps other than those stated in a claim. The article "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. The

15    invention may be implemented by means of hardware comprising several distinct elements, and by means of a suitably programmed computer. In the device claim enumerating several means, several of these means may be embodied by one and the same item of hardware. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage.

20